Maastricht University

Statistical inference for measures of predictive success

Citation for published version (APA):

Demuynck, T. (2014). *Statistical inference for measures of predictive success*. Maastricht University, Graduate School of Business and Economics. GSBE Research Memoranda No. 009
https://doi.org/10.26481/umagsb.2014009

Document status and date:
Published: 01/01/2014

DOI:
10.26481/umagsb.2014009

Document Version:
Publisher's PDF, also known as Version of record

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy
If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

Download date: 04 May. 2023

Maastricht University

Thomas Demuynck

Statistical inference for measures of predictive success

RM/14/009

GSBE

Maastricht University School of Business and Economics
Graduate School of Business and Economics

P.O Box 616
NL- 6200 MD Maastricht
The Netherlands

Statistical inference for measures of predictive success

Thomas Demuynck*

Abstract

We provide statistical inference for measures of predictive success. These measures are frequently used to evaluate and compare the performance of different models of individual and group decision making in experimental and revealed preference studies. We provide a brief illustration of our findings by comparing the predictive success of different revealed preference tests for models of intertemporal decision making.
JEL-codes: C10, C90, D12
Keywords: predictive success, revealed preference, experimental economics

1 Introduction

Given a behavioural model and an outcome space of possible observations, Selten (1991) distinguishes between three types of theories. A point theory gives a single element of the outcome space and predicts this point as the central tendency of the observations. A distribution theory gives a probability distribution over the outcomes space and predicts that observations are independently drawn according to this distribution. Finally, an area theory only predicts that the observed outcomes should lie in a certain subset of the outcome space. Given this classification, a distribution theory is more informative than either a point theory or an area theory.

Many applications in experimental and revealed preference settings, however, fall into the class of area theories. With respect to these theories, models are often evaluated on the basis of two metrics: the hit rate and the area. The hit rate gives the percentage of all observations that fall within the predicted subset of the outcome space. A high hit rate implies that many subjects have made choices that are consistent with the model's predictions. Hit rates, however, only capture one dimension of the model's performance. In general, the hit rate of a model will be higher if the model becomes less permissive (i.e. the model imposes weaker restrictions on the observed behaviour). Therefore, for an area theory to be meaningful it is desirable that the empirical test is sufficiently strong. The permissiveness can be measured by the 'area' of the test, which gives the relative size of the predicted subset compared to the set of all possible outcomes.[1]

*Maastricht University, Tongersestraat 53, 6711 LM Maastricht, Netherlands, email: t.demuynck@maastrichtuniversity.nl

[1]Of course, the 'size' of a set will always be conditional on a specific measure on the outcome space. Our framework will be flexible enough to allow for different specifications of this measure.

Generally, a favourable hit rate, for a specific behavioural model, provides convincing support for the model only if the associated area is sufficiently small. In practice, however, the two measures are almost always positively correlated, which in fact makes it interesting to define a summarizing measure that combines the two measures of empirical performance into a single metric, a so called measure of predictive success. Selten (1991) argues in favour of the functional specification that determines the predictive success as the difference between the hit rate and the area,

$$\text{predictive success} = \text{hit rate} - \text{area}.$$

This measure of predictive success is frequently used experimental studies[2] and has recently been advocated for use with revealed preference tests by Beatty and Crawford (2011).[3] In revealed preference studies, the area is usually quantified as one minus the Bronars (1987) power, which gives the probability that a randomly generated data sets (obtained from a uniform distribution on the budget hyperplanes) will fail the revealed preference test.

Although the predictive success measure is devoid of any statistical interpretation, the fact remains that its computation is based on the observed behaviour from a finite number of (randomly chosen) subjects. Considering the space of all possible observed behaviour as the relevant population, it becomes therefore possible to conduct statistical inference. In this note we use elementary large sample theory to construct asymptotically valid confidence intervals for various predictive success measures. Our results can be used to construct asymptotic valid hypothesis tests to verify whether the predictive success of a model is larger than some benchmark threshold or in order to compare the predictive success between different opposing models.

In the next section, we set out the framework and derive the statistical results. Section 3 contains an empirical illustration of our findings that compares the predictive success of different revealed preference tests for models of intertemporal decision making.

2 Framework

The building blocks of our framework are *data sets*, denoted by s. A data set may correspond to the outcome of an experiment for a single subject. In revealed preference theory, a data set usually consists of a finite set of price vectors $\{\mathbf{p}_t\}_{t\in T}$ together with corresponding consumption bundles $\{\mathbf{q}_t\}_{t\in T}$ describing the purchase behaviour of a single individual or household. We denote by Ω the set of all possible data sets that can be observed. An *experiment* is given by a finite number of data sets $\{s_i\}_{i\leq n}$ from Ω.

Hit rate An area theory for a certain model of behaviour predicts that the datasets will fall within a certain subset A of the outcome space Ω, e.g. in a revealed preference setting, we

[2]See among many others Huyck, Cook, and Battalio (1997), Hey (1998), Willinger and Ziegelmeyer (2001), Hey and Lee (2005), Gächter and Riedl (2006), Wang, Spezio, and Camerer (2010), Ehrhart, Gardner, von Hagen, and Keser (2007),Keser and Willinger (2007), Manzini, Mariotti, and Mittone (2010), Otto and Bolle (2011), Masatlioglu and Uler (2013).

[3]See, among others, Crawford (2010), Demuynck and Verriest (2013) and Deb, Gazzale, and Kotchen (2013) for applications.

could consider the subset of Ω that collects all data sets that satisfy the Generalized Axiom of Revealed Preference (Varian, 1982). Given such area theory, we consider the indicator function $I : \Omega \to \{0, 1\} : s \mapsto I(s)$ such that $I(s) = 1$ if and only if $s \in A$. The hit rate of the experiment $\{s_i\}_{i \leq n}$ is given by the proportion of datasets that fall within the set A.

$$r_n = \frac{1}{n} \sum_{i=1}^{n} I(s_i).$$

Area In order to define the area, we need a bit more work. To start, let us fix a data set $s_i \in \Omega$ and consider a probably space $(\Omega_i, \mathcal{B}_i, \mathbb{F}_i)$ which may depend on the specificities of the data set s_i. Here, $\Omega_i \subseteq \Omega$ is a subset of the outcome space. The set $\mathcal{B}_i$ is a sigma algebra on Ω_i such that the function $I(.)$ restricted to Ω_i is measurable and $\mathbb{F}_i : \mathcal{B}_i \to [0, 1]$ is a probability measure. We define the area of the dataset s_i by the function $\rho(s_i)$ where,

$$\rho(s_i) = \int I(s) \, \mathbb{F}_i(\mathrm{d}s).$$

Intuitively, $\rho(s_i)$ measures the size of the set A according to the measure $\mathbb{F}_i$. The area of the experiment $\{s_i\}_{i \leq n}$ is defined as the mean of the areas of the data sets in the experiment,

$$a_n = \frac{1}{n} \sum_{i=1}^{n} \rho(s_i)$$

In many experimental settings we have that Ω is finite, $\Omega_i = \Omega$ and $\mathbb{F}_i$ equals the uniform distribution on Ω, i.e. each individual data set is given an equal probability. In such setting, $\rho(s_i)$ will be the same for all s_i and the measure a_n will coincide with ρ. Typically, in a revealed preference setting, where $s_i = \{\mathbf{p}_t^i, \mathbf{q}_t^i\}_{t \in T}$, the measure $\mathbb{F}_i$ coincides with the probability law that randomly samples data sets $\tilde{s}_i = \{\mathbf{p}_t^i, \tilde{\mathbf{q}}_t^i\}_{t \in T}$ where $\tilde{\mathbf{q}}_t^i$ is obtained by a uniform draw from the hyperplane $\{\mathbf{q} \in \mathbb{R}_+^n | \mathbf{p}_t^i \mathbf{q} = \mathbf{p}_t^i \mathbf{q}_t^i\}$.[4] Observe, however, that our framework is flexible enough for other specifications of the probability measure $\mathbb{F}_i$.[5]

In some cases, it is possible to obtain $\rho(.)$ as a closed form solution. In other settings (like revealed preference theory) no closed form solutions are known. To encompass those situations, we allow $\rho(s_i)$ to be approximated by simulation. In such cases, we draw m i.i.d. data sets $\{\tilde{s}_1^i, \ldots, \tilde{s}_m^i\}$ using the probability measure $\mathbb{F}_i$ and compute the finite sample approximation,

$$\rho_m(s_i) = \frac{1}{m} \sum_{k=1}^{m} I(\tilde{s}_k^i).$$

The area of the experiment is then approximated by,

$$a_{n,m} = \frac{1}{n} \sum_{i=1}^{n} \rho_m(s_i),$$

Using the law of large numbers, we have that for $m \to \infty$, $a_{n,m} \to^P a_n$.

[4] This is analogue to the way the Bronars (1987) power is computed.

[5] See, for example, Andreoni, Gillen, and Harbaugh (2011) for such other measures.

Predictive success The hit rate r_n and the area $a_{n,m}$ can be combined in a measure of predictive success $p : [0,1]^2 \to \mathbb{R} : (r,a) \mapsto p(r,a)$. Intuitively, $p(r_n, a_{n,m})$ measures the performance of the behavioural model underlying the indicator function $I(.)$. Usually, p is increasing in its first argument and decreasing in its second. We assume that $p(.,.)$ is continuously differentiable.

Large sample results We consider the probability space $(\Omega, \mathcal{B}, \mathbb{P})$ where $\mathcal{B}$ is a sigma algebra on Ω and $\mathbb{P}$ is a probability distribution on Ω giving the law by which the individual data sets in the experiment are obtained. We assume that $\mathcal{B}$ is such that both the functions $I(.)$ and $\rho(.)$ are measurable.

The population hit rate and area are given by,

$$r = \int I(s)\mathbb{P}(\mathrm{d}s), \qquad \text{and} \qquad a = \int \rho(s)\mathbb{P}(\mathrm{d}s).$$

Consider an experiment $\{s_1, \ldots, s_n\}$ which is obtained from n i.i.d. draws according to the law $\mathbb{P}$. By the law of large numbers, we have that, as $n \to \infty$ and $m\, n^{-1} \to \infty$: $r_n \to^P r$ and $a_{n,m} \to^P a$. Further, using the classical central limit theorem, we have that,

$$\sqrt{n} \begin{pmatrix} r_n - r \\ a_{n,m} - a \end{pmatrix} \to N(0, \Sigma),$$

where,

$$\Sigma = \begin{bmatrix} r(1-r) & \int (I(s)-r)(\rho(s)-a)\mathbb{P}(\mathrm{d}s) \\ \int (I(s)-r)(\rho(s)-a)\mathbb{P}(\mathrm{d}s) & \int (\rho(s)-a)^2\mathbb{P}(\mathrm{d}s) \end{bmatrix},$$

is the asymptotic variance-covariance matrix. The elements of Σ can be consistently estimated by their finite sample analogues.

$$S_{n,m} = \begin{bmatrix} r_n(1-r_n) & \frac{1}{n}\sum_i (I(s_i)-r_n)(\rho_m(s_i)-a_{n,m}) \\ \frac{1}{n}\sum_i (I(s_i)-r_n)(\rho_m(s_i)-a_{n,m}) & \frac{1}{n}\sum (\rho_m(s_i)-a_{n,m})^2 \end{bmatrix}.$$

Using the continuous mapping theorem, we have that for $n \to \infty$ and $m\, n^{-1} \to \infty$: $p(r_n, a_{n,m}) \to^P p(r,a)$. Next, let d be the row vector of partial derivatives of the predictive success measure $p(r,a)$ evaluated at (r,a),

$$\delta = \begin{bmatrix} \frac{\partial p(r,a)}{\partial r} & \frac{\partial p(r,a)}{\partial a} \end{bmatrix}.$$

Using the delta method, we obtain that, for $n \to \infty$ and $m\, n^{-1} \to \infty$,

$$\sqrt{n}\left(p(r_n, a_{n,m}) - p(r,a)\right) \to N(0, \delta\Sigma\delta').$$

The variance, $\delta\Sigma\delta'$, can be consistently estimated by,

$$v_{n,m} = d_{n,m} S_{n,m} d'_{n,m},$$

where

$$d_{n,m} = \left[\frac{\partial p(r_n, a_{n,m})}{\partial r} \quad \frac{\partial p(r_n, a_{n,m})}{\partial a} \right].$$

If $\Phi(.)$ is the standard normal cdf function and,

$$\Phi(c_\alpha) - \Phi(-c_\alpha) = \alpha,$$

Then,

$$C_m^\alpha = \left[p(r_n, a_{n,m}) - c_\alpha \sqrt{\frac{v_{n,m}}{n}} \quad p(r_n, a_{n,m}) + c_\alpha \sqrt{\frac{v_{n,m}}{n}} \right],$$

is an asymptotic $\alpha \times 100\%$ confidence interval for the predictive success measure $p(r, a)$.

Comparing predictive success In many cases it is also interesting to compare two tests on the basis of their difference in predictive success. Consider two tests with hit rates and area equal to r, a and $\tilde{r}, \tilde{a}$, respectively. By the central limit theorem, we know that,

$$\sqrt{n} \begin{pmatrix} r_n - r \\ a_{n,m} - a \\ \tilde{r}_n - \tilde{r} \\ \tilde{a}_{n,m} - \tilde{a} \end{pmatrix} \to N\left(0, \Sigma_\Delta\right),$$

where Σ_Δ is the asymptotic variance covariance matrix whose elements can be consistently estimated using the finite sample plug ins. For example, the covariance between r and $\tilde{r}$ is equal to,

$$\int (I(s) - r)(\tilde{I}(s) - \tilde{r}) \mathbb{P}(\mathrm{d}s),$$

which can be consistently estimated by,

$$\frac{1}{n} \sum_i (I(s_i) - r_n)(\tilde{I}(s_i) - \tilde{r}_n).$$

We denote the estimator of the variance-covariance matrix by $S_{\Delta,n,m}$. Again, using the delta method, the asymptotic distribution of the difference in predictive success is given by,

$$\sqrt{n} \left[(p(r_n, a_{n,m}) - p(\tilde{r}_n, \tilde{a}_{n,m})) - (p(r, a) - p(\tilde{r}, \tilde{a})) \right] \to N\left(0, \delta_\Delta \Sigma_\Delta \delta_\Delta'\right).$$

Where δ_Δ is equal to the following row vector of partial derivatives,

$$\delta_\Delta = \left[\frac{\partial p(r,a)}{\partial r} \quad \frac{\partial p(r,a)}{\partial a} \quad -\frac{\partial p(\tilde{r},\tilde{a})}{\partial r} \quad -\frac{\partial p(\tilde{r},\tilde{a})}{\partial a} \right]$$

Set,

$$v_{\Delta,n,m} = d_{\Delta,n,m} S_{\Delta,n,m} d'_{\Delta,n,m}$$

where,

$$d_{\Delta,n,m} = \left[\frac{\partial p(r_n, a_{n,m})}{\partial r} \quad \frac{\partial p(r_n, a_{n,m})}{\partial a} \quad -\frac{\partial p(\tilde{r}_n, \tilde{a}_{n,m})}{\partial r} \quad -\frac{\partial p(\tilde{r}_n, \tilde{a}_{n,m})}{\partial a} \right]$$

Then

$$\left[p(r_n, a_{n,m}) - p(\tilde{r}_n, \tilde{a}_{n,m}) - c_\alpha \sqrt{\frac{v_{\Delta,n,m}}{n}} \quad p(r_n, a_{n,m}) - p(\tilde{r}_n, \tilde{a}_{n,m}) + c_\alpha \sqrt{\frac{v_{\Delta,n,m}}{n}} \right],$$

is an asymptotic $\alpha \times 100\%$ CI for $p(r, a) - p(\tilde{r}, \tilde{a})$.

3 Illustration

We illustrate our results using various revealed preference tests for different models of intertemporal decision making. The first model is the standard life cycle (LC) model where an individual optimizes a time separable additive utility function $\sum_t \delta^t u(\mathbf{q}_t)$ subject to an intertemporal budget constraint $\mathbf{p}_t \mathbf{q}_t + a_t = I_t + (1 + r_t) a_{t-1}$. Here $\delta < 1$ is a subjective discount rate, $\mathbf{p}_t$ are the period t prices, a_t is the value of assets at period t, I_t is the contemporaneous income and r_t is the interest rate. Data sets for this model are determined by prices, quantities and interest rates for a finite number of periods, $s = \{\mathbf{p}_t, \mathbf{q}_t, r_t\}_{t=1,\ldots|T|}$. The revealed preference conditions for this life cycle model were derived by Browning (1989).

For the second model, let us first single out a habit forming good c. The habits (H) model replaces the intertemporal separable utility function by a utility function of the form $\sum_t \delta^t u(\mathbf{q}_t, c_{t-1})$. Here, the consumption of the addictive good in period $t-1$ is allowed to influence the utility in period t. The revealed preference characterization of this model was given by Crawford (2010).

Our third model, the habits as durables (HAD) model, considers a variant where the intertemporal utility function is given by $\sum_t \delta^t u(\mathbf{q}_t, A_t)$ and where $A_t = \beta A_{t-1} + c_t$, represents a stock of addiction with depreciation rate β that determines how fast the addiction wears off. This is the rational addiction model put forward by Becker and Murphy (1988). The revealed preference characterization of this model was derived by Demuynck and Verriest (2013).

As a final fourth model, we consider the static utility maximization model where the household maximizes each period a time independent utility function $u(\mathbf{q})$ subject to a budget constraint $\mathbf{p}_t \mathbf{q} = m_t$ for some level of expenditure m_t. The revealed preference conditions for this model is given by the Generalized Axiom of Revealed Preference (Varian, 1982).

In addition to the four models, we consider 3 measures of predictive success.

$$\begin{aligned} p_1(r, a) &= r - a, \\ p_2(r, a) &= \frac{r}{a}, \\ p_3(r, a) &= \frac{r - a}{1 - a}. \end{aligned}$$

The first measure takes the difference between the hit rate and the area and is the measure that has become standard in the literature. It is bounded between -1 and 1. In the best case scenario, $r \to 1$ and $a \to 0$. This gives a predictive success close to one. In such case, most data sets pass

the test while the area is very small. In the worst case scenario, $r \to 0$ and $a \to 1$, which gives a predictive success close to minus one. In this case, almost all observations are inconsistent with the model while the area is almost equal to the outcome space Ω. In intermediate cases, the measure of predictive success is found somewhere between minus one and plus one. Zero is a natural benchmark where $r = a$.

The second measure takes the ratio of the hit rate and area. Intuitively, p_2 measures the density of the observed data sets within the predicted area. It is bounded from below by zero. The natural benchmark, where $r = a$, gives a predictive success equal to one. The third measure is obtained from the first measure by dividing it by the maximal value that it can obtain for fixed a. It can also be written as $1 - \frac{1-r}{1-a}$. Intuitively, this predictive success measure will be higher, the lower the density outside the predicted area. Its benchmark is equal to zero. We refer to Selten (1991) for a more thorough discussion of the differences between these predictive success measures.

Data description We use data from the Encuesta Continua de Presupuestos Familiares. This data set contains detailed information on consumed quantities and prices for a large sample of Spanish households. We refer to Browning and Collado (2001), Crawford (2010) and Demuynck and Verriest (2013) for a more detailed explanation of this data set. The observations range from 1985 until 1997 and are obtained on a quarterly basis. Every quarter, new households are participating in the moving panel and others are dropped. There are a maximum of eight consecutive observations per household. We consider 14 nondurable commodity categories,[6] and take tobacco as the habit forming good.[7] Finally, we simulate the areas $\rho(s_i)$ using 1000 random draws per data set (in other words, we set m equal to 1000).

Results Table 1 provides the results on the estimates of $p(r, a)$ for the different measures and the 95% asymptotic confidence intervals. For the first measure, the highest estimate is for the HAD model which is also the only model whose confidence interval excludes the benchmark value 0. For the second measure, the highest value is found for the LC model. However, this model also has the highest variance, which makes its value highly uncertain. Both H and HAD models exclude 1 from the 95% confidence intervals. The last measure gives qualitatively similar results as the first measure.

[Table 1 about here.]

Table 2 gives the mean values and asymptotic confidence intervals for the difference in predictive success between the different revealed preference tests. Many intervals include the value of zero meaning that the hypothesis of equal predictive success cannot be rejected at the 5% level.

[6] In particular, we have (1) Food and non-alcoholic drinks at home, (2) Alcohol, (3) Tobacco, (4) Energy at home, (5) Services at home, (6) Nondurables at home, (7) Nondurable medicines, (8) Medical services, (9) Transportation, (10) Petrol, (11) Leisure, (12) Personal services, (13) Personal non–durables, (14) Restaurants and bars.

[7] We further restrict the sample to the subset of households for which the wife is outside of the labour market and for which we have observations for all eight quarters. We further restrict the sample to households which have strict positive consumption for the addictive good in all periods. This procedure leaves a sample of 671 households.

Exceptions to this are the differences between the GARP and the HAD test for measures 1 and 2, the difference between the LC and H test for measure 1 and the difference between the H and HAD test for all predictive success measures under consideration.

[Table 2 about here.]

References

Andreoni, J., Gillen, B. J., Harbaugh, W. T., 2011. The power of revealed preference tests: Ex-post evaluation of experimental design. Tech. rep.

Beatty, T. K. M., Crawford, I. A., 2011. How demanding is the revealed preference approach to demand. American Economic Review 101, 2782–2795.

Becker, G. S., Murphy, K. M., 1988. A theory of rational addiction. Journal of Political Economy 96, 675–700.

Bronars, S. G., 1987. The power of nonparametric tests of preference maximization. Econometrica 55, 693–698.

Browning, M., 1989. A nonparametric test of the life-cycle rational expectations hypothesis. International Economic Review 30, 979–992.

Browning, M., Collado, M. D., 2001. The response of expenditures to anticipated income changes: Panel data estimates. American Economic Review 91, 681–692.

Crawford, I., 2010. Habits revealed. Review of Economic Studies 77, 1382–1402.

Deb, R., Gazzale, R. S., Kotchen, M. J., 2013. Testing motives for charitable giving. a revealed preference methodology with experimental evidence. Tech. rep.

Demuynck, T., Verriest, E., 2013. I'll never forget my first cigarette: A revealed preference analysis of the habits as durables model. International Economic Review 54, 717–738.

Ehrhart, K.-M., Gardner, R., von Hagen, J., Keser, C., 2007. Budget processes. theory and experimental evidence. Games and Economic Behavior 59, 279–295.

Gächter, S., Riedl, A., 2006. Dividing justly in bargaining problems with claims. Social Choice and Welfare 27, 571–594.

Hey, J. D., 1998. An application of Selten's measure of predictive success. Mathematical Social Sciences 35, 1–15.

Hey, J. D., Lee, J., 2005. Do subjects separate (or are they sophisticated)? Experimental Economics 8, 233–265.

Huyck, J. B. V., Cook, J. P., Battalio, R. C., 1997. Adaptive behavior and coordination failure. Journal of Economic Behavior and Organization 32, 483–503.

Keser, C., Willinger, M., 2007. Theories of behaviour in principal-agent relationships with hidden actions. European Economic Review 51, 1514–1533.

Manzini, P., Mariotti, M., Mittone, L., 2010. Choosing monetary sequences: theory and experimental evidence. Theory and Decision 69, 327–354.

Masatlioglu, Y., Uler, N., 2013. Understanding the reference effect. Games and Economic Behavior 82, 403–423.

Otto, P. E., Bolle, F., 2011. Matching markets with price bargaining. Experimental Economics 14, 322–348.

Selten, R., 1991. Properties of a measure of predictive success. Mathematical Social Sciences 21, 153–167.

Varian, H., 1982. The nonparametric approach to demand analysis. Econometrica 50, 945–974.

Wang, J. T., Spezio, M., Camerer, C. F., 2010. Pinocchio's pupil. using eyetracking and pupil dilation to underunder truth telling and deception in sender-receiver games. American Economic Review 100, 984–1007.

Willinger, M., Ziegelmeyer, A., 2001. Strength of the social dilemma in a public goods experiment: An exploration of the error hypothesis. Experimental Economics 4, 131–144.

Table 1: Mean values and 95% confidence intervals for the predictive success measures

	GARP	LC	H	HAD
p_1	0.0123	0.0013	0.0332	0.1505
	$[-0.0084 \quad 0.0330]$	$[-0.0016 \quad 0.0042]$	$[0.0000 \quad 0.0664]$	$[0.1126 \quad 0.1884]$
p_2	1.0136	6.6666	1.1470	1.4006
	$[0.9907 \quad 1.0365]$	$[-6.4608 \quad 19.7941]$	$[1.0001 \quad 1.2940]$	$[1.2996 \quad 1.5016]$
p_3	0.1306	0.0013	0.0429	0.2410
	$[-0.0891 \quad 0.3503]$	$[-0.0017 \quad 0.0042]$	$[0.0000 \quad 0.0859]$	$[0.1803 \quad 0.3016]$

Table 2: Mean and 95% confidence intervals for difference in predictive success

	GARP - LC	GARP - H	GARP - HAD
p_1	0.0110	-0.0209	-0.1382
	$[-0.0098 \quad 0.0319]$	$[-0.0581 \quad 0.0162]$	$[-0.1803 \quad -0.0959]$
p_2	-5.6530	-0.1334	-0.3870
	$[-18.7802 \quad 7.4740]$	$[-0.2797 \quad 0.0128]$	$[-0.4894 \quad -0.2846]$
p_3	0.1293	0.0877	-0.1104
	$[-0.0903 \quad 0.3490]$	$[-0.1315 \quad 0.3068]$	$[-0.3351 \quad 0.1143]$
	LC-H	LC-HAD	H-HAD
p_1	-0.0320	-0.1492	-0.1172
	$[-0.0651 \quad 0.0012]$	$[-0.1871 \quad -0.1113]$	$[-0.1518 \quad -0.0827]$
p_2	5.5196	5.2661	-0.2536
	$[-7.5993 \quad 18.6386]$	$[-7.8576 \quad 18.3897]$	$[-0.3798 \quad -0.1273]$
p_3	-0.0417	-0.2397	-0.1980
	$[-0.0845 \quad 0.0011]$	$[-0.3003 \quad -0.1791]$	$[-0.2504 \quad -0.1457]$